纵向汉明距离判别法：对潜在属性的发展追踪[*]

刘耀辉[1] 陈琦鹏[1] 徐慧颖[1] 詹沛达[1,2**]

([1]浙江师范大学教师教育学院心理系，金华 321004；

[2]浙江省智能教育技术与应用重点实验室，金华 321004)

摘　要 研究通过在纵向诊断数据分析中引入计算简单、耗时少的汉明距离判别法(HDD)，提出了纵向HDD (Long-HDD)。与 HDD 相比，Long-HDD 额外使用汉明距离刻画个体在相邻时间点上对属性掌握的相依性，以利用前一时间点信息提高当前时间点的分类准确性。三个模拟研究的结果主要表明：在分析纵向诊断数据时，与参数化模型相比，Long-HDD 的分类准确性几乎不受样本量影响，在样本量较小时表现更优；且其计算耗时更少，更有利于提供及时性诊断反馈。实证研究结果表明 Long-HDD 可用于分析实践测评数据，且其追踪诊断结果与参数化模型的存在一致性。

关键词 认知诊断；非参数分类法；纵向数据分析；汉明距离

1 引言

追踪学生知识和技能的变化，不仅有利于评估教学方法的有效性，还可以绘制出学生的发展轨迹，有利于优化教学过程和实施补救教学(詹沛达等, 2021)。对学生发展轨迹的追踪依赖于多次测量所收集的纵向数据。近年来，纵向认知/学习诊断，即评估学生的潜在属性(如, 知识和技能)并确定其在一个时期内优势与不足，逐渐受到研究者和教育者的关注(e.g., Zhan, 2020b; 王立君等, 2020)。学者们提出了不同的纵向诊断分类模型(diagnostic classification models, DCMs)为纵向诊断数据分析提供方法支持(见 Zhan, 2020b; 詹沛达等, 2021)，如，基于高阶潜在结构模型的纵向 DCM 和基于潜在转换分析的纵向 DCM。这些研究结果表明，参数化纵向 DCMs 可以有效诊断学生个体或群体的学习成长轨迹。

然而，参数化模型的实际应用可能会遇到一些障碍，特别是在纵向诊断最适用的小规模教育测评(如, 班级或学校水平的测评)中。例如，专家判定的认知假设(如, 连接或分离缩合规则)与学生实际应用的存在偏差，进而选用了基于错误认知假设的 DCM；需要大样本以保证参数估计和结果分类的准确性(Chiu & Köhn, 2015)；在应用相对复杂的模型时，分析人员也需要一定的编程能力。此外，参数化模型通常需要较多的计算时间，特别是对于一些使用马尔科夫链蒙特卡洛(MCMC)算法的纵向 DCMs。这些障碍无疑提高了一线教育工作者应用参数化模型的难度，不利于其在实践教学环境中的应用和推广。

[*] 本研究得到浙江省哲学社会科学规划“之江青年理论与调研专项课题”(22JQN38YB)和教育部人文社会科学青年基金项目(19YJC190025)资助。

[**] 通讯作者: 詹沛达, E-mail: pdzhan@gmail.com

为避免上述问题，研究者们提出了一些非参数分类(nonparametric classification, NPC) 法(Chiu, Douglas, et al., 2009; Chiu & Douglas, 2013; Chiu, et al., 2018; Wang & Douglas, 2015; 康春花等, 2019)。与参数化 DCM 相比，NPC 法因不需要进行参数估计，所以它对样本量没有明确要求；即 NPC 法的应用摆脱了对大样本量的依赖，即使样本量为 1 也可以使用。此外，NPC 法不需要考虑参数化模型可能存在的参数估计不收敛问题，并且在各种认知假设下均有良好的表现(Chiu & Douglas, 2013; Wang & Douglas, 2015; Chiu et al., 2018; Akbay, 2016)。同时，由于 NPC 法计算简单，其对应用者的编程水平要求相对较低，并且可以实现对被试的快速分类，便于提供及时性诊断反馈。

然而，目前尚未有研究试图使用 NPC 法来分析纵向认知诊断数据；即 NPC 法在追踪学生发展方面的心理测量学表现还不明晰。随着“为学习而测评”理念的普及(詹沛达等, 2021)，考虑到实践教学对纵向认知诊断的需求以及 NPC 法的简便性，将两者结合并探究 NPC 法在追踪学生发展方面的心理测量学表现是值得尝试的。为此，本研究的主要目的是提出一种纵向 NPC 法，以期实现对个体属性掌握情况变化的追踪。

首先，回顾两种相关方法：一个是有代表性的参数化纵向 DCM——纵向高阶 DINA (Long-DINA) 模型(Zhan, Jiao, Liao, & Li, 2019)；另一个是应用于横断诊断数据的汉明距离判别法(Hamming distance discrimination, HDD) (Chiu & Douglas, 2013)。其次，介绍了拟提出的纵向 NPC 法——纵向 HDD (longitudinal HDD, Long-HDD)。接着通过模拟研究评估新方法在纵向诊断测评数据分析中的表现，并与横断 HDD 和 Long-DINA 模型进行对比研究。然后，以一则实证数据分析为例呈现新方法的实践可应用性。最后，总结该研究的一些发现并讨论未来研究方向。

2 相关方法简介

2.1 Long-DINA 模型

已有研究表明，Long-DINA 模型在纵向认知诊断数据分析中具有良好的心理测量学表现(Zhan, 2020b; Zhan, Jiao, Liao, & Li, 2019)。该模型可表示为：

一阶：

$$P(y_{nit}=1\,|\,\boldsymbol{\alpha}_{nt},g_{it},s_{it})=g_{it}+(1-g_{it}-s_{it})\prod_{k=1}^{K}\alpha_{nkt}^{q_{ikt}}\,, \tag{1}$$

二阶：

$$P(\alpha_{nkt}=1\,|\,\theta_{nt},\beta_k,\delta_k)=\frac{\exp(\beta_k\theta_{nt}-\delta_k)}{1+\exp(\beta_k\theta_{nt}-\delta_k)}, \tag{2}$$

三阶：

$$\boldsymbol{\theta}_n=(\theta_{n1},\ldots,\theta_{nT})'\sim MVN_T(\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{3}$$

其中，g_{it} 和 s_{it} 分别表示时间点 t 上题目 i 上的猜测和失误参数；$\boldsymbol{\alpha}_{nt} = (\alpha_{n1t}, ..., \alpha_{nKt})'$表示时间点 t 上被试 n 的属性掌握向量，$\alpha_{nkt} \in \{0, 1\}$; q_{ikt} 表示时间点 t 上题目 i 是否考查属性 k，$q_{ikt} \in \{0, 1\}$；θ_{nt} 为时间点 t 上被试 n 的高阶能力；β_k 和 δ_k 分别表示所有时间点上属性 k 的斜率和难度参数；$\boldsymbol{\mu} = (\mu_1, ..., \mu_T)'$为多元正态分布的均值向量，$\boldsymbol{\Sigma}$ 是该分布下的协方差矩阵，

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & \\ \vdots & \ddots & \\ \sigma_{1T} & \cdots & \sigma_T^2 \end{bmatrix}, \tag{4}$$

其中 σ_{1T} 表示时间点 1 的和时间点 T 的高阶能力之间的协方差。另外，作为纵向测验的起点及后续时间点的参考点，θ_{n1} 被约束为服从标准正态分布。

2.2　HDD 法

如未明确说明，文中“HDD”仅指代横断 HDD。HDD 是一种易于理解且常见的 NPC 法，它比较观察作答向量(ORP)和理想作答向量(IRP)之间的汉明距离。其中，汉明距离是两个相同长度的向量在相应位置上不同元素的个数：

$$d_{hm}(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{I} |A_i - B_i|, \tag{5}$$

其中$|A_i - B_i| = 1$ 表示向量 **A** 中第 i 个元素与向量 **B** 中的第 i 个元素不相等，反之$|A_i - B_i| = 0$ 则表示相等。例如，(1,1,0,1)和(1,0,0,1)之间的汉明距离为 1。

HDD 使用与 ORP 距离最近的 IRP 所对应的属性向量作为学生的属性掌握向量(AMP) (Chiu & Douglas, 2013; 罗照盛等, 2015)。HDD 中，学生 n 的 ORP 与第 c (c=1, 2, …, 2^K)个 IRP 之间的汉明距离可以表示为：

$$d_{hm}(ORP_n, IRP_c) = \sum_{i=1}^{I} |y_{ni} - y_{ci}^*|, \tag{6}$$

其中，y_{ni} 表示学生 n 在题目 i 上的作答结果，y^*_{ci} 为第 c 个 IRP 的第 i 个元素。当仅有一个 IRP 与 ORP 的距离最短时，与这个 IRP 所对应的属性向量则被视为该学生的 AMP；而当有多个 IRP 与 ORP 的距离最短时，则可以通过 R 方法(随机抽取法)、B 方法(贝叶斯方法)和 W 方法(加权汉明距离法)[1](对应图 1 中的矩形虚线部分)，筛选出其认为最有可能的 AMP (Chiu & Douglas，2013; 罗照盛等，2105; 康春花等，2017)。

需要注意的是 HDD 也可用于分析纵向诊断数据，这与使用横断参数模型分析纵向诊断数据的做法类似(Zhan, 2020a)。具体而言，使用 HDD 分析纵向诊断数据时，有两种计算策略：一种是重复使用 HDD 分别分析不同时间点的数据(即独立计算策略的 HDD，记为 HDD-D)，另一种是将所有时间点数据整合为一个数据并使用 HDD 进行一次性分析(即同时性计算策略的 HDD，记为 HDD-T)。理论上两种计算

[1] 由于有研究表明该三种方法在大多数情况下的判准率几乎无差异，因此，不在此做详细介绍，仅选用简明的 R 方法(随机抽取法)来处理“一对多”的情况，详情可查阅康春花等(2017)。

策略的结果相同；基于模拟研究的数据分析结果也表明两者的表现几乎一样且 HDD-D 的计算耗时更少（理论说明及数据分析结果见附录）。为精简，下文仅使用 HDD-D。

3 Long-HDD 方法的提出

通常，纵向研究中，在不同时间点上指代同一潜在特质的多个潜在变量(如, 公式 2 中指代同一潜在特质——高阶能力——的多个潜在变量 θ_{nt})之间是高相关的。在 Long-DINA 中，这种相关性可由多元正态分布来表示(见公式 3)。同样，在 NPC 法中，如何在不涉及模型参数的情况下，处理相邻时间点间 AMP 的相关性是一个亟需解决的问题。

在对虚拟测评过程性数据的分析中(刘耀辉等, 2022)，有研究者使用了距离判别的方法来处理不同行为序列之间的关联性，如汉明距离(Bergner et al., 2014)和编辑距离(Hao et al., 2015)。在解决问题时，不同个体的动作序列之间距离越近，表明他们的动作序列越一致、解决问题能力越接近。换句话说，动作序列之间距离越近，它们之间相关性就越高，越倾向于被分为同一类。受此启发，在提出的纵向 NPC 法中，汉明距离将被再次用来表示同一学生在不同时间点上 AMP 之间的关联性。本研究假设同一学生在相邻时间点上的 AMP 是高度相关的，因此该学生在两个相邻时间点上 AMP 之间的汉明距离也应是最小的。

Long-HDD 用汉明距离来表示同一学生相邻时间点上对属性掌握情况的相关性或依赖性，以利用前一个时间点的信息来提高当前时间点的分类精度。Long-HDD 中，同一学生相邻时间点上 AMP 之间的相关性或依赖性可表示为：

$$d_{hm}(AMP_{nt}, \boldsymbol{\alpha}^{*}_{nc(t+1)}) = \sum_{k=1}^{K}\left|\alpha_{nkt} - \alpha^{*}_{nck(t+1)}\right|, \quad (7)$$

其中，AMP_{nt}为被试 n 在时间点 t 上的 AMP, $\boldsymbol{\alpha}^{*}_{nc(t+1)}$ 为时间点 t+1 上与学生 n 的 ORP 具有最小汉明距离的 IRP 集合所对应的属性模式集合中的第 c 个属性向量，$\alpha^{*}_{nck(t+1)}$ 则为该第 c 个属性向量中的第 k 个属性的状态(0 或 1)。α_{nkt} 为时间点 t 上学生 n 对属性 k 的掌握状态(即时间点 t 上学生 n 的 AMP 中的第 k 个属性)。

如图 1 所示，在某一特定的时间点内，Long-HDD 与 HDD 在计算方式上是一致的。两者的主要区别是在判断学生 n 在时间点 2 的 AMP (即 AMP_{n2})时，须额外计算 IRP 集合 2 中每个 IRP 所对应的属性向量与该学生在时间点 1 的 AMP (即 AMP_{n1})之间的汉明距离(见公式 7)；然后使具有最小汉明距离的属性向量作为该学生时间点 2 上的 AMP[2]。总的来说，该方法通过多次使用汉明距离来联接学生在相邻时间点上 AMP 之间的关联性。当 $T = 1$ 或在时间点 1 上，Long-HDD 等同于 HDD。

2 理论上，在比较 AMPn_1 和时间 2 中的“对应属性模式 2”的汉明距离时，仍有可能出现“一对多”的情况。因考虑到出现该种情况的概率比较小，本文并未对该过程中可能出现“一对多”的情况进行进一步探究。本文中的处理方式为: 不管是否会在该次汉明距离比较中出现“一对多”的情况，依然采用随机的方式。也就是说，从该次汉明距离比较得出的结果(与 AMPn_1 的汉明距离最短的属性模式)中随机出一个作为 AMPn_2(如果该汉明距离比较的结果中只有一个属性模式与

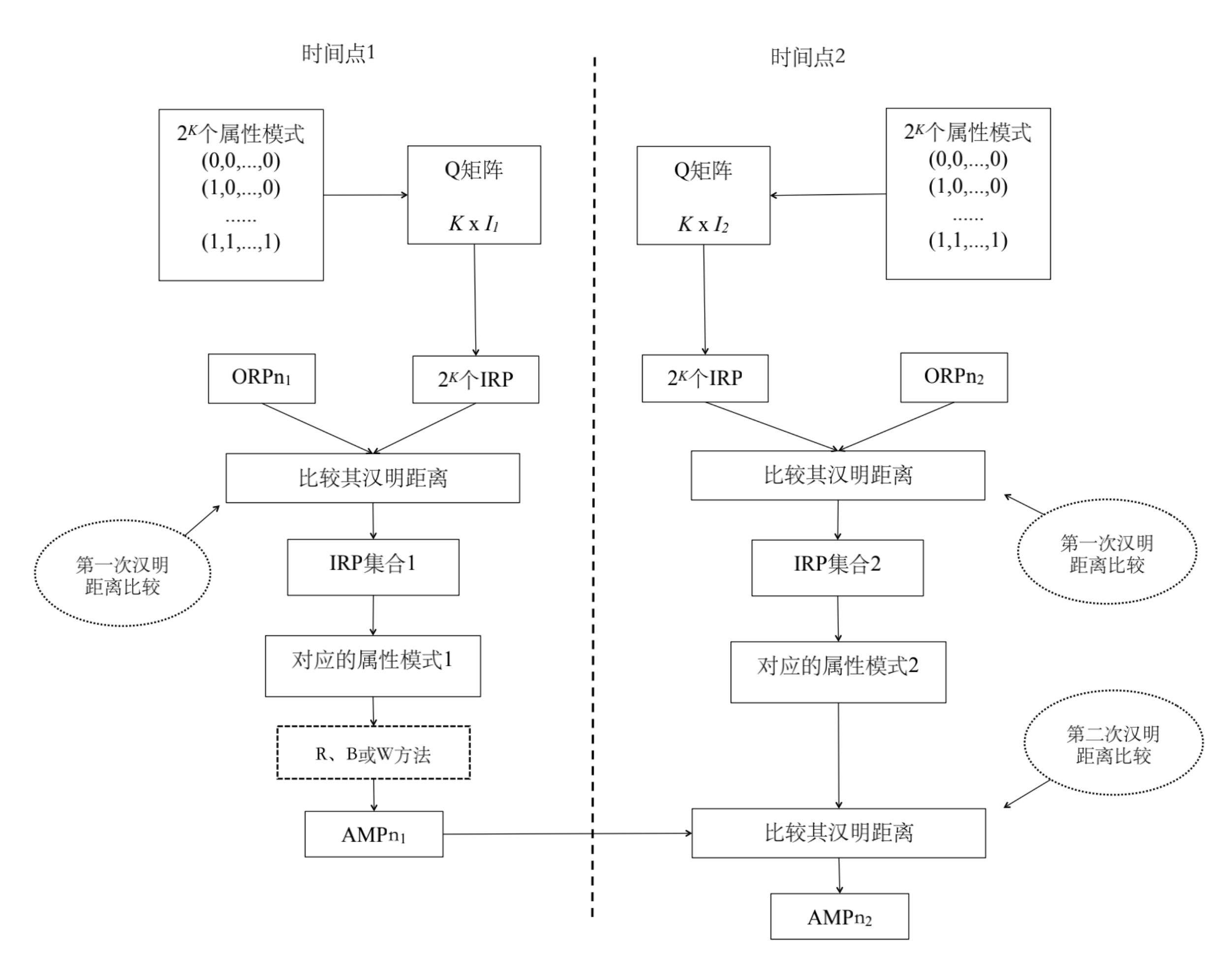


图 1　Long-HDD 应用流程图.

4 模拟研究

本文开展三个模拟研究来探究 Long-HDD 在纵向诊断数据分析中的表现，即探究在追踪学生发展时它是否能提供准确的分类结果；并且比较 Long-HDD、HDD 和 Long-DINA 模型的分类准确性。

4.1　模拟研究 1

4.1.1　研究设计及数据生成

本部分研究包含 5 个自变量：(1)考虑到 NPC 法的适用情境，本研究侧重于小规模测评，样本量 $N=$ 25、50、100 和 300；(2)每个时间点上的题目数量 $I=25$ 和 50；(3)每个时间点上所考查的属性数量 $K=$ 3 和 5；(4)测试时间点数量 $T=2$ 和 3；(5)数据分析方法 $M=$ Long-HDD、HDD 和 Long-DINA。

AMPn_1 距离最小，则该属性模式则为 AMPn_2)

在每个时间点 Q 矩阵中，将构成可达/单位矩阵的前 K 道题目设定为锚题，并在所有时间点固定不变(Zhan, Jiao, Liao, & Li, 2019)。除锚题外，其他题目所考察的属性向量是以相同概率从包含≥2 个属性的属性向量中随机提取。考虑到实际测验中，题目的猜测和失误概率可能呈现负相关(Zhan, Jiao, Liao, & Bian, 2019)，在每个时间点上题目 i 的猜测参数 g_{it} 和失误参数 s_{it} 由二元正态分布产生：

$$\begin{pmatrix} logit(g_{it}) \\ logit(s_{it}) \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\begin{pmatrix} -2.197 \\ -2.197 \end{pmatrix}, \begin{pmatrix} 1 & -0.6 \\ -0.6 & 1 \end{pmatrix}\right). \quad (8)$$

该设定下所有题目的猜测和失误概率服从正偏态分布(平均值≈0.1，最小值≈0.01，最大值≈0.6)。各时间点上锚题的猜测和失误概率均被固定为 0.1。

为保证方法之间对比的公平性，本研究未直接采用 Long-HDD 或 Long-DINA 模型作为数据生成模型。被试的属性掌握增长按如下方法生成：在第一个时间点，每个学生的真实 AMP 是以相同的概率从所有可能的 2^K 个属性向量中随机抽取。在随后的时间点，每个学生的真实 AMP 的生成参考了 Li 等(2016)的研究。具体而言，设定所有属性的相邻时间点之间的转移概率相等：

$$\mathrm{P}(\alpha_{nkt} \to \alpha_{nk(t+1)}) = \begin{cases} \mathrm{P}(0 \to 0) = 0.8 \\ \mathrm{P}(0 \to 1) = 0.2 \\ \mathrm{P}(1 \to 0) = 0.05 \\ \mathrm{P}(1 \to 1) = 0.95 \end{cases}. \quad (9)$$

最后，根据生成的真实 AMP、题目参数(异常反应概率)和 Q 矩阵，分别在各时间点上使用 DINA 模型生成观察数据。每种模拟条件下生成 50 批数据。

4.1.2 分析

每个模拟条件下均使用 Long-HDD、HDD-D 和 Long-DINA 分析数据。对 HDD-D 而言，重复使用 HDD 分析每个时间点上的数据。对 Long-HDD 而言，对每个时间点的数据逐一进行分析，即基于 HDD，结合前一个时间点的分类结果，辅以判定当前时间点上学生的 AMP。对 Long-DINA 模型而言，采用同时性参数估计策略(Zhan, 2020a)，这涉及到将多个时间点的作答数据重新整合为一个作答矩阵，然后将其作为一个整体进行分析。使用 MCMC 算法实现对 Long-DINA 模型的参数估计：包含两条链，每条链采样 8,000 次，前 5,000 次用于预热；根据潜在量尺缩减因子(PSRF)<1.1 的标准，待估计参数均已收敛。

使用属性向量正确分类率(PCCR)和纵向 PCCR (LPCCR) (Zhan, Jiao, Liao, & Li, 2019)评估各方法的分类准确性：$\mathrm{PCCR}_t = \sum_{r=1}^{R}\sum_{n=1}^{N}\frac{I[\hat{\boldsymbol{\alpha}}_{ntr}=\boldsymbol{\alpha}_{nt}]}{NR}$，$\mathrm{LPCCR} = \sum_{r=1}^{R}\sum_{n=1}^{N} I[\forall t, \hat{\boldsymbol{\alpha}}_{ntr} = \boldsymbol{\alpha}_{nt}]/NR$，其中， $I[\cdot]$是一个指示函数，N 是样本量，R 是迭代次数，t 为第 t 个时间点，$\hat{\boldsymbol{\alpha}}_{nt}$和$\boldsymbol{\alpha}_{nt}$分别表示学生 n 在时间点 t 的估计 AMP 和真实 AMP。

4.1.3 结果

图2呈现了三种方法在不同模拟条件下各时间点的 PCCR 和 LPCCR。首先，与已有研究的发现类似(Chiu & Douglas, 2013; Zhan, Jiao, Liao, & Li, 2019; 康春花等, 2017)，随着所测属性数量的增加和测验时间点数量的增加，三种方法的分类准确性均有所下降。相比之下，随着题目数量的增加，三种方法的分

类准确性均有所上升。其次，样本量对 HDD-D 和 Long-HDD 的影响较小。且在所有条件下，Long-HDD 的分类准确性均高于 HDD-D 的，表明在分析纵向数据时，考虑相邻时间点 AMP 之间的相关性或依赖性是必要的。然后，当 Long-HDD 和 Long-DINA 进行比较式时，前者的相对优势在样本量较小的时候更为明显，而后者在样本量增加至 100 以上时显示出与前者匹配的分类准确性，且后者的相对优势随样本量的继续增加而增加。

4.2 模拟研究 2

4.2.1 研究设计、数据生成和分析

研究 2 为进一步考察在小样本条件下 Long-HDD 在不同大小的属性转移概率下的表现。更高的属性转移概率意味着有更大比例的学生对属性的掌握状态发生变化。本研究包含 4 自变量：(1) 同一属性在相邻时刻间的转移概率 $P(0 \to 1) = 0.5$ (中等水平)和 $P(0 \to 1) = 0.8$ (高水平)，见公式(10); (2) 每个时间点上的题目数量 $I = 25$ 和 50；(3) 每个时间点上所考查的属性数量 $K = 3$ 和 5；(4) 数据分析方法 $M =$ Long-HDD、HDD-D 和 Long-DINA。另外，T=3 和 N=25；Q 矩阵和题目参数的设置、观察数据的生成，以及分析方法同研究一保持一致。

$$\text{中等水平增长：} P(\alpha_{nkt} \to \alpha_{nk(t+1)}) = \begin{cases} P(0 \to 0) = 0.5 \\ P(0 \to 1) = 0.5 \\ P(1 \to 0) = 0.05 \\ P(1 \to 1) = 0.95 \end{cases},$$

$$\text{高水平增长：} P(\alpha_{nkt} \to \alpha_{nk(t+1)}) = \begin{cases} P(0 \to 0) = 0.2 \\ P(0 \to 1) = 0.8 \\ P(1 \to 0) = 0.05 \\ P(1 \to 1) = 0.95 \end{cases}. \quad (10)$$

4.2.2 结果

图 3 呈现了三种方法在不同模拟条件下的 PCCR 和 LPCCR。结果表明三种方法的判准率的相对优劣在中等水平增长和高水平增长条件下相对稳定；其中，Long-HDD 的判准率仍然是最高的。结合研究 1 和研究 2 的结果表明，在小样本条件下，Long-HDD 的判准率优于 HDD-D 和 Long-DINA 的，且该优势几乎不受出现掌握状态变化的学生的比例的影响。

4.3 模拟研究 3

4.3.1 研究设计、数据生成和分析

本研究拟在小样本条件下采用 Long-DINA 生成数据；此时，若 Long-HDD 的表现仍优于或不输于 Long-DINA 的，则可进一步凸显在小样本测验情境下使用 Long-HDD 的相对优势。对此，本研究仅包含 1 个自变量：属性数量 $K = 3$ 和 5；并将样本量、时间点数和每个时间点上题目数量分别固定为 N=25、$T = 3$ 和 I=25，Q 矩阵和题目参数的生成方式与研究 1 保持一致。另外，参照詹沛达等人(2021)的研究，

在 Long-DINA 二阶模型（公式 2）中，固定属性难度参数 $\boldsymbol{\beta} = (-1, -0.5, 0, 0.5, -1)'$，属性区分度参数 ξ_k 固定为 1.5；在三阶模型（公式 3）中，相邻时刻间一般潜在能力 θ_n 的均值变化设定为 0.5，量尺变化(即标准差增加的倍数)设定为$\sqrt{1.25}$。作答结果数据依据 $y_{nit} \sim \text{Bernoulli}(P(y_{nit} = 1))$ 生成，其中 $P(y_{nit} = 1)$为公式 1。分析方法同研究 1 保持一致。

4.3.2 结果

图 4 呈现了三种方法在不同模拟条件下的 PCCR 和 LPCCR。结果显示三种方法的判准率差异较小；这表明即便在相对不公平的对比条件下（使用 Long-DINA 模型作为数据生成模型），Long-HDD 方法也能提供与数据生成模型相当的判准率。另外，三种方法对 50 组数据的平均计算耗时和排序分别为：$K = 3$ 时 Long-DINA (138.95 秒) > Long-HDD (0.11 秒) > HDD-D (0.07 秒)；$K = 5$ 时 Long-DINA (174.78 秒) > Long-HDD (0.44 秒) > HDD-D (0.29 秒)。显然，属性数量会影响参数估计时间；另外，非参数方法的计算耗时要远小于参数化模型的；并且，由于 Long-HDD 方法针对“一对多”的情况需要额外计算汉明距离，所以其计算耗时略多于 HDD-D。表 1 呈现了 50 组数据中 HDD-D 中出现“一对多”的比例。可以看到 $K = 3$ 时各时间点中“一对多”的平均比例约 5%左右；$K = 5$ 时各时间点中“一对多”的平均比例约 10%左右。即属性数量的增加会导致“一对多”比例的增加。但需要强调的是 HDD-D 中出现“一对多”的比例受题目参数的影响（罗照盛等, 2015），而本研究中题目参数的生成是基于二元正态分布随机生成的（即无法固定题目参数的影响），因此该比例结果仅供参考。

综上所述，模拟研究结果表明(1)Long-HDD 在纵向诊断数据分析中具有较高的分类准确性；(2)Long-HDD 的表现几乎不受样本量的影响，在小样本情况下表现优于 Long-DINA；(3)Long-HDD 的表现几乎不受出现掌握状态变化的学生的比例的影响；(4)小样本条件下，即便以 Long-DINA 作为数据生成模型，Long-HDD 的表现也不输于 Long-DINA 的；和(5)Long-HDD 的计算耗时低于 Long-DINA 的。

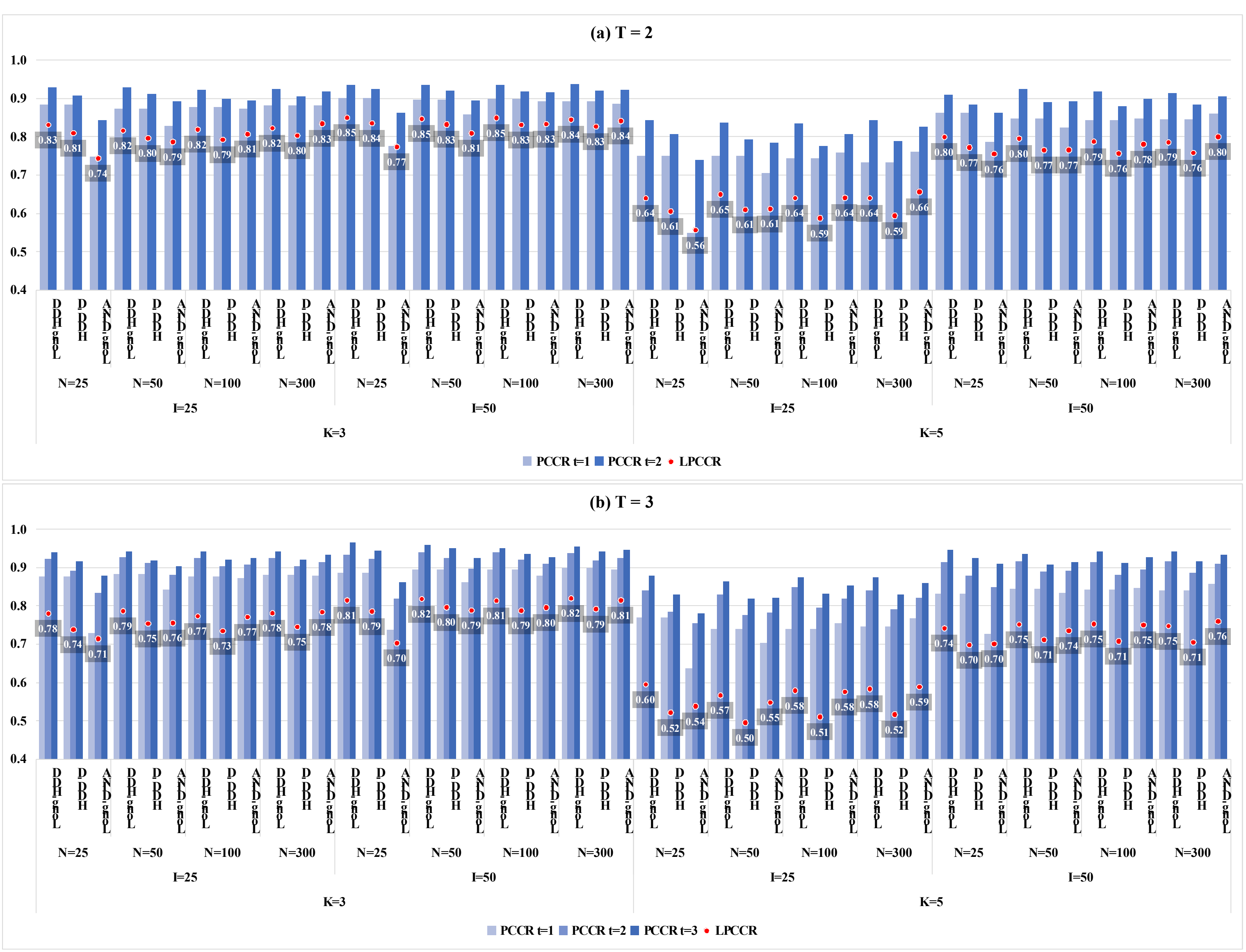


图 2　模拟研究 1 中 Long-HDD, HDD-D 和 Long-DINA 的分类准确率.

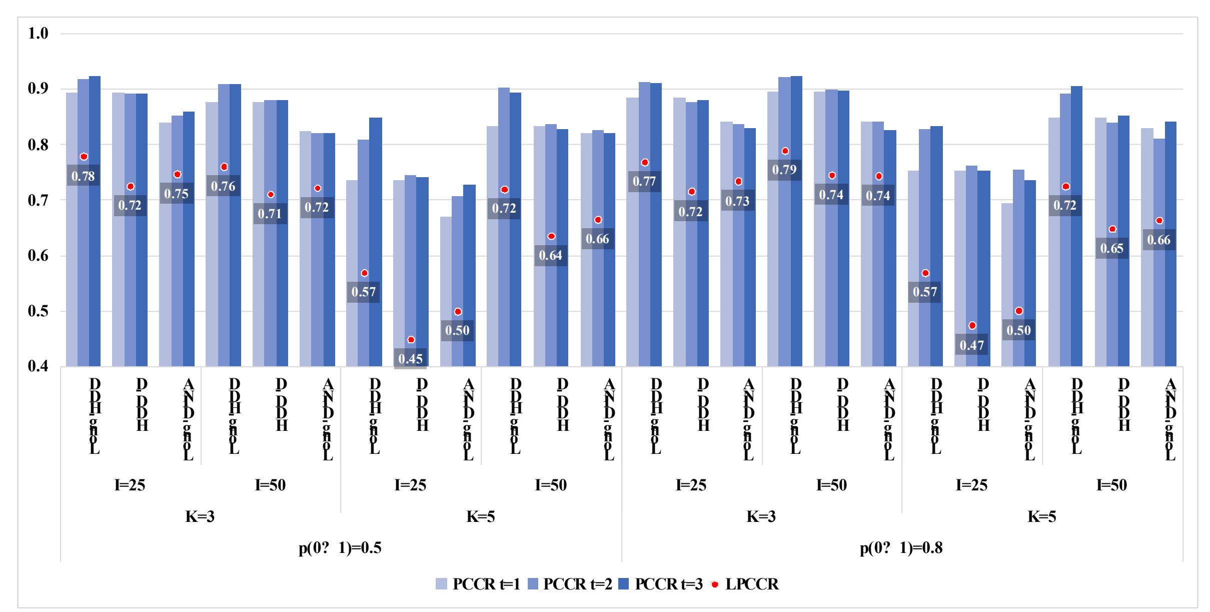


图 3　模拟研究 2 中 Long-HDD, HDD-D 和 Long-DINA 的分类准确率.

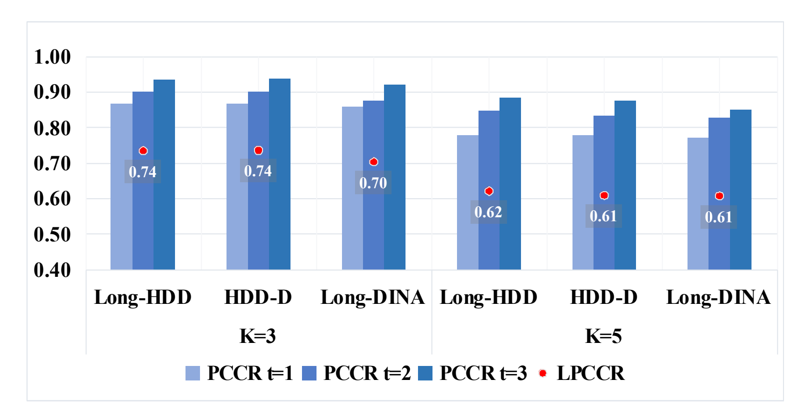


图 4　模拟研究 3 中 Long-HDD, HDD-D 和 Long-DINA 的分类准确率.

表 1　模拟研究 3 中 HDD-D 方法中不同时刻出现“一对多”的比例.

K	“一对多”比例	$t=1$	$t=2$	$t=3$
3	最大值	0.16	0.20	0.12
	最小值	0.00	0.00	0.00
	平均值	0.05	0.06	0.03
5	最大值	0.24	0.28	0.20
	最小值	0.00	0.00	0.00
	平均值	0.13	0.12	0.09

注：最大值、最小值和平均值分别为 50 组数据中的最大、最小和平均的“比例”。

5　实证数据分析

5.1　数据说明

实证数据来源于某中学七年级 90 名学生的数学测验(Tang & Zhan, 2021)。为避免记忆/练习效应，该纵向测验包括三套测试题目不同的平行测试；分三次施测，相邻两次施测时间间隔为一周。三套平行测试具有相同的 Q 矩阵(见图 5)，每次测试包含 18 道题目，共测量 6 个属性。分别使用 Long-HDD、

HDD-D 和 Long-DINA 模型分析该数据，并计算每个施测时间点下，每种方法下学生在不同属性上的掌握占比(即掌握人数占总人数的比例)。

5.2 结果

图 6 呈现了三种方法下六个属性的掌握占比随时间发展趋势。首先，整体而言，学生对各属性的掌握占比均有提升，但不同属性有不同的变化趋势。其次，不同方法对属性变化趋势的追踪略有差异。具体而言，对属性 1、3 和 4，三种方法的追踪结果较为一致；而对于属性 2、5 和 6，Long-HDD 与 HDD-D 的追踪结果更一致，两者与 Long-DINA 模型的追踪结果存在一定差异。另外，在分析该实证数据时，HDD-D 方法中出现“一对多”的比例在三个时间点上分别为 0.88，0.72 和 0.48；而 Long-HDD 方法中第二次计算最小汉明距离时出现“一对多”的比例为 0 和 0.11 (第二次计算最小汉明距离对应图 1 中比较 AMPn_1 和“对应的属性掌握模式 2”的汉明距离)。这不仅表明了 Long-HDD 在实际的数据分析中存在很大的应用空间，而且也说明再次使用最小汉明距离可以有效缩小 t (t >1)时刻学生可能的 AMP 范围，获得较为稳定的判准结果。但遗憾的是，由于非参数法无法像参数化模型一样提供模型-数据拟合指标值，因此，我们并无法对比三种方法对该数据的拟合或匹配程度。总之，实证研究表明 Long-HDD 可用于分析实践测评数据且 Long-HDD 的追踪诊断结果与 Long-DINA 模型的存在一定的一致性。

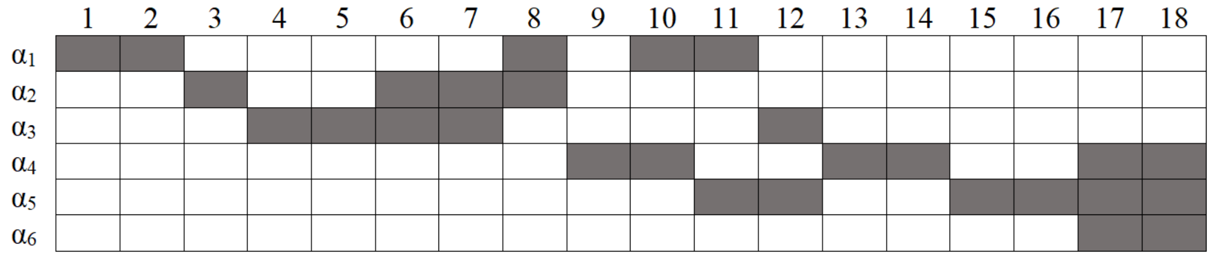
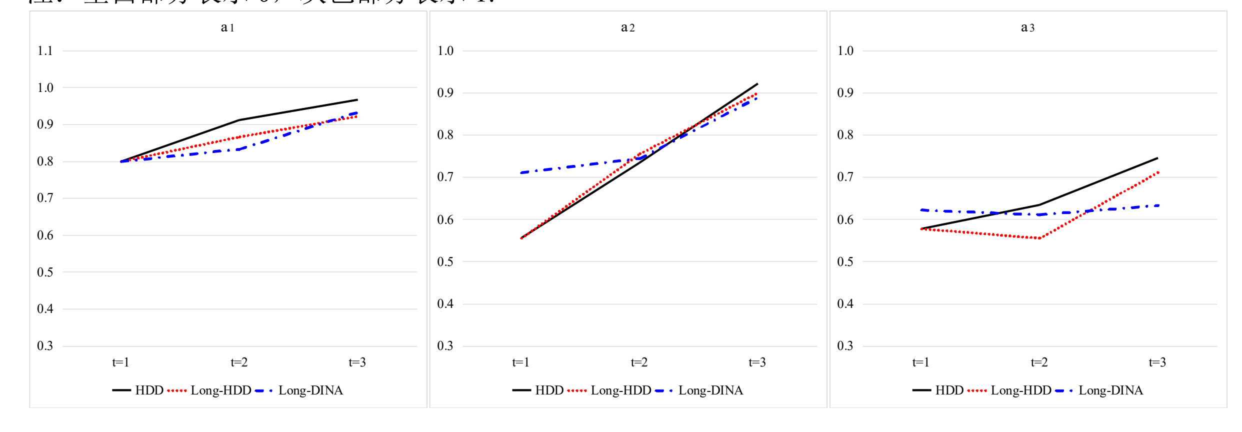


图 5 有理数测验的 Q 矩阵

注：空白部分表示 0，灰色部分表示 1.



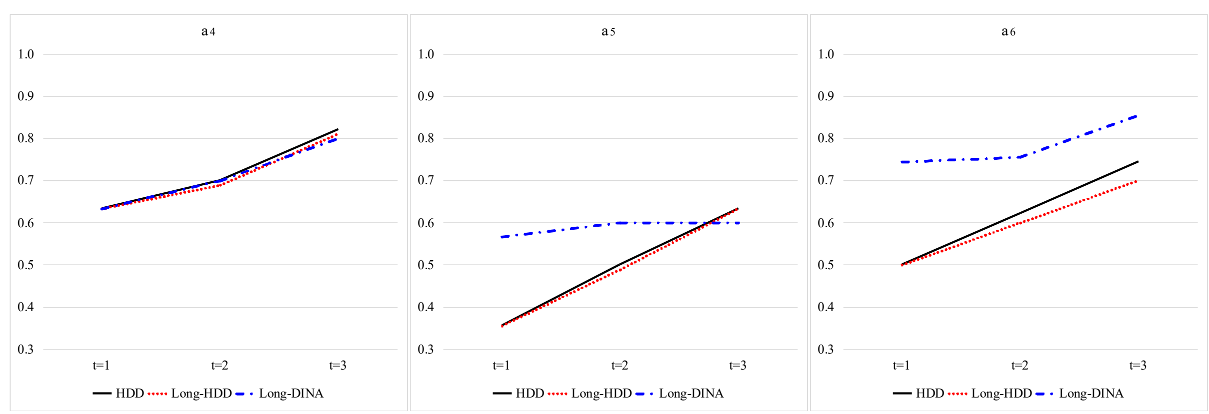


图 6 六个属性的掌握占比随时间发展趋势.

6 讨论、展望与研究结论

6.1 讨论与展望

本研究通过在纵向诊断数据分析中引入计算简单、耗时少的 HDD，提出了 Long-HDD，以期促进纵向认知诊断在实践教学中的应用。与 HDD 相比，Long-HDD 使用汉明距离刻画个体在相邻时间点上对属性掌握情况的相关性或相依性，以利用前一个时间点的信息来提高当前时间点的分类准确性。该方法不仅降低了纵向诊断数据分析的难度，还拓宽了 NPC 法的应用范围。模拟研究结果主要表明 Long-HDD 可用于分析纵向诊断数据，也能刻画学生的潜在属性变化；且由于其计算简单耗时少，更有利于实现即时性诊断反馈，这在班级和学校层面的小规模测评中尤为重要。实证研究表明 Long-HDD 可用于分析实践测评数据且其追踪诊断结果与 Long-DINA 模型的存在一定的一致性。另外，在纵向数据的分析中，Long-HDD 通过多次汉明距离的比较，可获得较为稳定的判准结果，相对于采用随机化方法的 HDD 来说，有更高的诊断分类一致性。

作为一种非参数方法，Long-HDD 缺少相关的数据拟合指标，即无法评估其对数据的拟合程度。实际上，在班级水平小规模测评中，诊断结果的准确性并不是实践者关注的首要问题。即便不进行诊断性测评，任课教师也可以根据学生的原始作答来推断每一个学生对知识点的掌握情况。从该角度看，基于方法的诊断结果更多地是为了降低老师工作量，为一对一有针对性干预提供参考性建议；而至于老师是否要依据诊断结果为学生提供干预，最终仍由老师决定。反之，在实践应用中如果样本量足够(如, 大于 300 人)且无计算耗时压力，仍推荐使用可提供模型-数据拟合程度的参数化模型。

由于能力和精力有限，本文仍有一些局限性值得后续做进一步探究。首先，Long-HDD 是 HDD 的拓展，因此，HDD 的理论局限性也存在于 Long-HDD 中。例如，本文中 HDD 只考虑了最简单的汉明距离。在实践中，不同的属性可能有不同的权重，后续研究中可以考虑将加权 HDD (Chiu & Douglas, 2013)纳入 Long-HDD 中。其次，本研究假设属性遵循连接缩合规则，即学生只有在掌握题目所考察的所有属性时，才会有较高的正确作答概率。而这一假设可能在一定程度上制约了 Long-HDD 的实践应用。未来，可尝试将广义 NPC 法 (Chiu et al., 2018)纳入 Long-HDD 中，以提高新方法在不同认知假设下的适用性。然后，本研究中每个属性在相邻时间点之间的转移概率被

设定为相等。尽管作者认为该设置对本研究的结论没有影响，但 Long-HDD 在可变的转移概率下的具体表现仍值得进一步探索。然后，理论上 Long-HDD 仅在时间点 t (t >1)的 IRP 集合中包含多个 IRP 时(即存在“一对多”)才起作用。而当 IRP 集合中只有一个 IRP 时，Long-HDD 的分类结果等同于 HDD 的。Wang 和 Douglas (2015)认为增加题目数量会减少“一对多”情况的发生，然而，本研究结果发现，即便题目数量从 25 增加到 50，似乎并没有减少 Long-HDD 相对于 HDD 的优势。罗照盛等人(2015)探究了在不同属性层级结构和作答失误概率对出现“一对多”的影响；其研究发现随着作答失误概率的增加，在任一属性层级结构中会出现更多的“一对多”。此外，康春花等人(2017)研究表明 Q 矩阵中所含可达矩阵的数量也会影响 HDD 的分类准确性。因此，作者认为除题目数量外，“一对多”情况的发生还受到 Q 矩阵的复杂性、题目质量及属性层级等因素的影响，而具体的影响程度也值得后续做进一步探究。

6.2 研究结论

Long-HDD 在纵向诊断数据分析中可以提供较高的分类精度，且由于不受样本量的影响、计算简单、耗时少，其相对于参数化纵向 DCM (如, Long-DINA)，更有利于提供即时性诊断反馈，更适用于如班级和学校层面的小规模纵向测评。

参考文献

康春花，杨亚坤，曾平飞. (2017). 海明距离判别法分类准确率的影响因素. *江西师范大学学报(自然科学版).41*(04), 394-400.

康春花，杨亚坤，曾平飞. (2019). 一种混合计分的非参数认知诊断方法: 曼哈顿距离判别法. *心理科学,42*(2), 455-462.

刘耀辉, 徐慧颖, 陈琦鹏, 詹沛达. (2022). 基于过程数据的问题解决能力测量及数据分析方法. *心理科学进展, 30*(3), 522-535.

罗照盛，李喻骏，喻晓锋，高椿雷，彭亚风. (2015). 一种基于 Q 矩阵理论朴素的认知诊断方法. *心理学报*(02), 264-272.

王立君, 唐芳, 詹沛达. (2020). 基于认知诊断测评的个性化补救教学效果分析：以"一元一次方程"为例. *心理科学, 43*(06), 1490-1497.

詹沛达, 潘艳方, 李菲茗. (2021). 面向"为学习而测评"的纵向认知诊断模型. *心理科学, 44*(1), 214-222.

Akbay, L. (2016). Relative efficiency of the nonparametric approach on attribute classification for small sample cases. *Journal Of European Education*, 6(1).

Bergner, Y., Shu, Z., & von Davier, A. A. (2014). Visualization and confirmatory clustering of sequence data from a simulation-based assessment task. *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 177–184).

Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*(2), 225-250.

Chiu, C. Y., Douglas, J. A., & Li, X. D. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

Chiu, C.Y. & Köhn, H. F. (2015). A general proof of consistency of heuristic classification for cognitive diagnosis models. *British Journal of Mathematical and Statistical Psychology, 68*(3), 387–409.

Chiu, C. Y., Sun, Y. & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika 83*, 355–375.

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining, 7*(1), 33–50.

Tang, F., & Zhan, P. (2021). Does diagnostic feedback promote learning? Evidence from a longitudinal cognitive diagnostic assessment. *AERA Open*, *7*.

Wang, S. Y., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika, 80*(1), 85–100.

Zhan, P. (2020a). A Markov estimation strategy for longitudinal learning diagnosis: providing timely diagnostic feedback. *Educational and Psychological Measurement, 80*(6), 1145–1167.

Zhan, P. (2020b). Longitudinal learning diagnosis: minireview and future research directions. *Frontiers in Psychology*, *11*, 1185.

Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics, 44*(3), 251-281.

Zhan, P., Jiao, H., Liao, M., & Bian, Y. (2019). Bayesian DINA modeling incorporating within-item characteristic dependency. *Applied Psychological Measurement, 43*(2), 143-158.

Longitudinal Hamming Distance Discrimination: Developmental Tracking of Latent Attributes

Liu Yaohui[1], Chen Qipeng[1], Xu Huiying[1], Zhan Peida[1,2]

([1]Department of Psychology, College of Teacher Education, Zhejiang Normal University;
[2]Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua 321004, China)

Abstract

Longitudinal cognitive diagnostics can assess students' strengths and weaknesses over time, profile students' developmental trajectories, and can be used to evaluate the effectiveness of teaching methods and optimize the teaching process. Researchers have proposed different longitudinal diagnostic classification models, which provide methodological support for the analysis of longitudinal cognitive diagnostic data. Although these parametric longitudinal cognitive diagnostic models can effectively assess students' growth trajectories, their requirements for coding ability and sample size hinder their application among frontline educators, and they are time-consuming and not conducive to providing timely feedback. On the one hand, the nonparametric approach is easy to calculate, efficient to apply, and provides timely feedback; on the other hand, it is free from the dependence on sample size and is particularly suitable for analyzing assessment data at the classroom or school level. Therefore, this study attempts to apply nonparametric method to longitudinal cognitive diagnostic assessments for tracking student's learning trajectories.

This study extended the longitudinal Hamming distance discriminant (Long-HDD) based on the Hamming distance discriminant (HDD), which uses the Hamming distance to represent the dependence between attribute mastery patterns of the same student at adjacent time points. To explore the performance of Long-HDD in longitudinal cognitive diagnostic data, we conducted three simulation studies and an empirical study and compared the classification accuracy of the HDD, Long-HDD, and Long-DINA models. The purpose of simulation study 1 was to compare the performance of performance of three methods under different simulation conditions. Simulation study 2 focused on the classification accuracy of the three methods at moderate attributes transfer probability level ($p(0\rightarrow1)=0.5$, $p(1\rightarrow0)=0.05$) and high attributes transfer probability level ($p(0\rightarrow1)=0.8$, $p(1\rightarrow0)=0.05$). To further highlight the advantages of the Long-HDD in small-scale assessments, the Long-DINA model was used as the data generation model in Study 3. At this point, if Long-HDD still outperforms or does not lose out to Long-DINA model's, the relative advantage of using Long-HDD in a small-scale assessments can be further highlighted. Furthermore, an empirical study was conducted to illustrate the application of the Long-HDD.

Under the comparison of the three methods, the results of the simulation studies showed that (1) Long-HDD had higher classification accuracy in longitudinal diagnostic data analysis; (2) Long-HDD performed almost independently of sample size and performed better with a smaller sample size compared to Long-DINA; and (3) Long-HDD consumed much less computational time than Long-DINA. In addition, the results of the empirical study showed that there was good consistency between the results of the

Long-HDD and the Long-DINA model in tracking changes in attribute development. The percentage of mastery of each attribute increased with the increase of time points.

In summary, the long-HDD proposed in this study extends the application of nonparametric methods to longitudinal cognitive diagnostic data and can provide high classification accuracy. Compared with parameterized longitudinal DCM, it can provide timely diagnostic feedback due to the fact that it is not affected by sample size, simple calculation, and less time-consuming. It is more suitable for small-scale longitudinal assessments such as class and school level.

Keywords: cognitive diagnosis; nonparametric classification; longitudinal data analysis; Hamming distance

附 录

S1 两种计算策略 HDD 结果的一致性

假设有两个时间点，每个时间点用 I 个题目测量 K 个属性。首先，解释一些下文会用到的缩写：ORP_1、ORP_2、AMP_1、AMP_2、IRP 集、IRP 集合 1、IRP 集合 2 分别代表学生在时间点 1 的观察作答向量、时间点 2 的观察作答向量、时间点 1 的属性掌握向量、时间点 2 下的属性掌握向量、筛选出的与 ORP 距离最短的 IRP 的集合、在时间点 1 与 ORP_1 距离最短的 IRP 集合和在时间点 2 与 ORP_2 距离最短的 IRP 集合。

S1.1 采用独立计算策略的 HDD-D

这种情况下，在每个时间点可得到 2^K 个 IRP，每个 IRP 的长度等于 I。通过比较每个时间点的 ORP 和 IRP 的汉明距离。可以得到 IRP 集合 1 和 IRP 集合 2。假设从 IRP 集合 1 中随机选择的 $IRPr_1$ 和 ORP_1 之间的汉明距离为 s_1，从 IRP 集合 2 中随机选择的 $IRPr_2$ 和 ORP_2 之间的汉明距离为 s_2，那么(ORP_1，ORP_2)和(IRP_1，IRP_2)之间的总汉明距离为 s_1+s_2。例如，ORP_1 为(0,1,1)，ORP_2 为(0,1,0)，IRP_1 是(1,1,1)，IRP_2 是(1,1,1)，那么(ORP1, ORP2) (即(0,1,1,0,1,0))和(IRP1, IRP2) (即(1,1,1,1,1,1))的汉明距离为 3，且等于 ORP_1 和 IRP_1 之间的距离 s_1 加上 ORP_2 和 IRP_2 之间的距离 s_2。如图 S1 所示，如果 IRP 集 1 有 x 个元素，IRP 集 2 有 m 个元素，那么最后判准的结果为某个(AMP_1, AMP_2)的概率是 $1/xm$。

S1.2 采用同时性计算策略的 HDD-T

在这种情况下，纵向数据被作为一个整体数据进行分析，Q 矩阵的表示就略有不同。假设每个时间点的测试中有 5 个题目，3 个属性，两个时间点，其 Q 矩阵可表示为图 S2。$I_1, I_2, \ldots I_5$ 是在第一个时间点测量的题目，$I_6, I_7, \ldots I_{10}$ 是在第二个时间点测量的题目。K_{11}、K_{21}、K_{31} 代表在第一时间点测量的三个属性，K_{12}、K_{22}、K_{32} 代表在第二时间点测量的同样三个属性。

图 S3 显示了在纵向数据中使用 HDD 与同时性计算策略的流程。通过两个时间点，可以得到 2^{2K} 个属性向量，从而再得到 2^{2K} 个 IRP。同样，可得到一个 IRP 集合，其中的 IRP 与(ORP_1, OPR_2)保持最小的汉明距离。因为汉明距离计算的是两个等长的字符串之间对应位上不同元素的数量。ORP 和 IRP 在不同时间下的汉明距离是相互独立的。不难得到 IRPr 和(ORP1，ORP2)之间的距离是 s_1+s_2，等于($y^*_1, y^*_2, \ldots y^*_I$)和($ORP_1$)之间的距离加上($y^*_{I+1}, \ldots y^*_{2I}$)和($ORP_2$)的距离。IRPr 是从 IRP 集合中随机选择的，该集合等于 IRP 集合 1 和 IRP 集合 2 的所有组合(在图 S1 中)(即 $w = xm$)。最后，判准结果为某个(AMP_1，AMP_2)的概率也是 $1/xm$。

S1.3 小规模模拟研究

为进一步研究 HDD-D 和 HDD-T 在数据分析中是否也存在一致性，我们基于模拟研究 3 中的条件，对比了 HDD-D 和 HDD-T 的属性判准率和计算耗时。表 S1 呈现了 HDD-D 和 HDD-T 的 PCCR 和 LPCCR，发现两者之间的差异很小(约 0.5%)。但两者的计算耗时具有较大差异，当 $K = 3$ 时，HDD-D 和 HDD-T 的平均计算耗时分别为 0.07 秒和 4.25 秒；当 $K = 5$ 时，HDD-D 和 HDD-T 的平均计算耗时分别为 0.29 秒和 232.37 秒（超过了 Long-DINA 模型的 174.78 秒）。表明两者的计算

时间都受到属性数量的影响，但 HDD-T 所受的影响更大。其主要原因在于 $T = 3$ 和 $K = 5$ 条件下 HDD-T 中有 2^{15}=32768 种属性模式需要进行距离计算。总之，数据分析结果表明 HDD-D 与 HDD-T 的属性判准率几乎一样，且前者计算耗时更少。

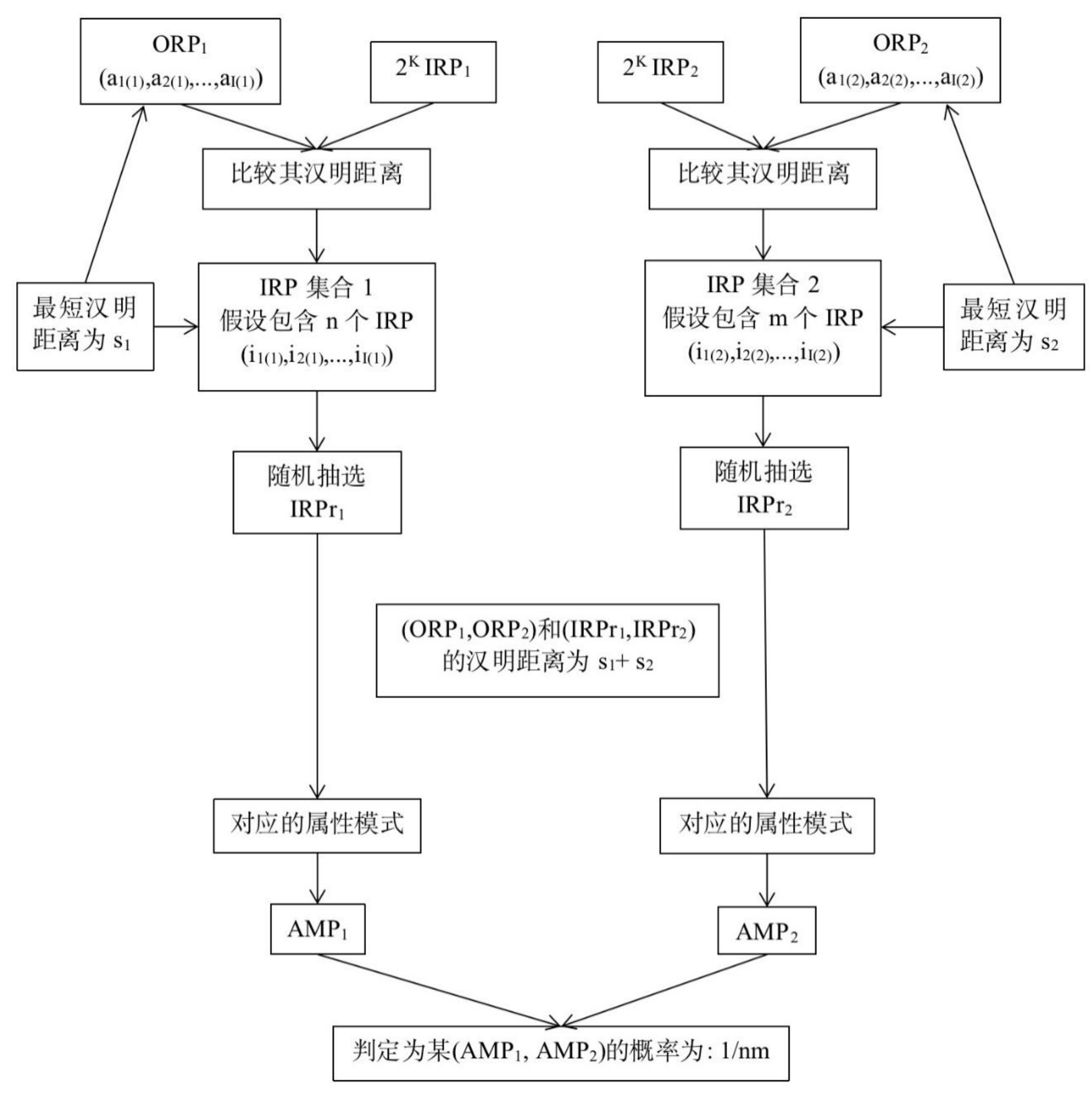


图 S1　独立计算策略 HDD-D 示意图

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
α_{11}	■			■						
α_{21}		■			■					
α_{31}			■	■	■					
α_{12}						■				■
α_{22}							■		■	■
α_{32}								■	■	

图 S2 同时性计算策略 HDD 的纵向 Q 矩阵

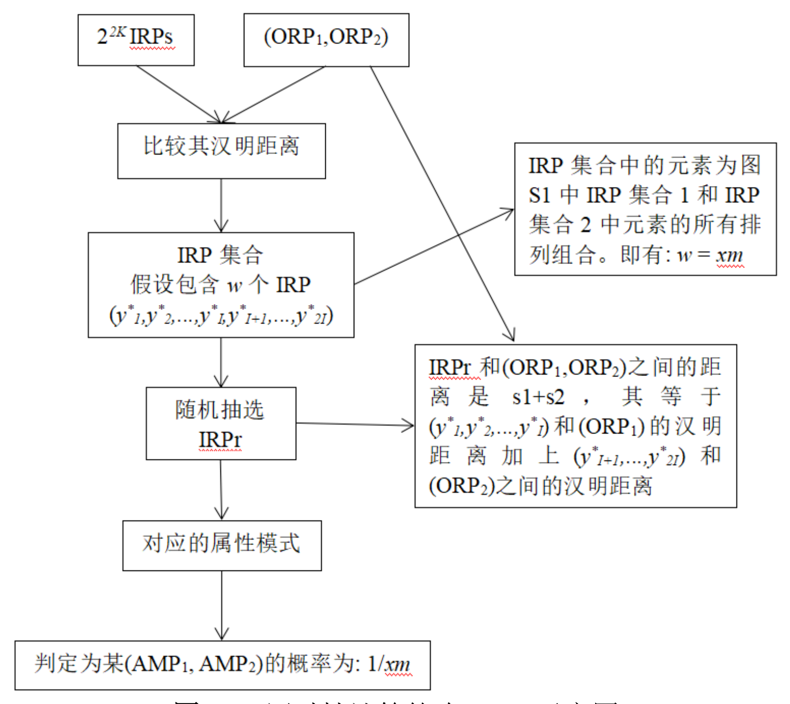


图 S3 同时性计算策略 HDD 示意图

表 S1 模拟研究 3 条件下 HDD-D 与 HDD-T 的属性判准率.

K	HDD 计算策略	PCCR			LPCCR
		$t = 1$	$t = 2$	$t = 3$	
3	HDD-D	0.867	0.900	0.937	0.737
	HDD-T	0.870	0.893	0.936	0.731
5	HDD-D	0.779	0.834	0.876	0.615
	HDD-T	0.784	0.838	0.879	0.619